What is machine learning ?

Warning!

Does every observable natural phenomenon follow a law?

Math language

Law as a function

How would we *know* that function?

Origin of approximation theory

Weierstrass Theorem (1885)

Given $f : [a, b] \to \mathbb{R}$ continuous and an arbitrary $\varepsilon > 0$, there exists an algebraic polynomial p such that

$$|f(x) - p(x)| \leq \varepsilon, \quad \forall x \in [a, b] \subset \mathbb{R}.$$

Stone WeierstrassTheorem (Weierstrass 1887, simplified proof Stone 1948)

Which is a polynomial ?

1. x^3
2. x^{π^e}
3. $\sum_{i=0}^{n} \frac{x^i}{i!} \; n \in \mathbb{N}$
4. $\sum_{i=0}^{\infty} \frac{x^i}{i!}$

Formalizing the approach

- Data come from $\mathcal{P}(\mathcal{D})$.
- Condition: Independent identically distributed iid
- The information is the characteristic vector $\mathcal{X}$

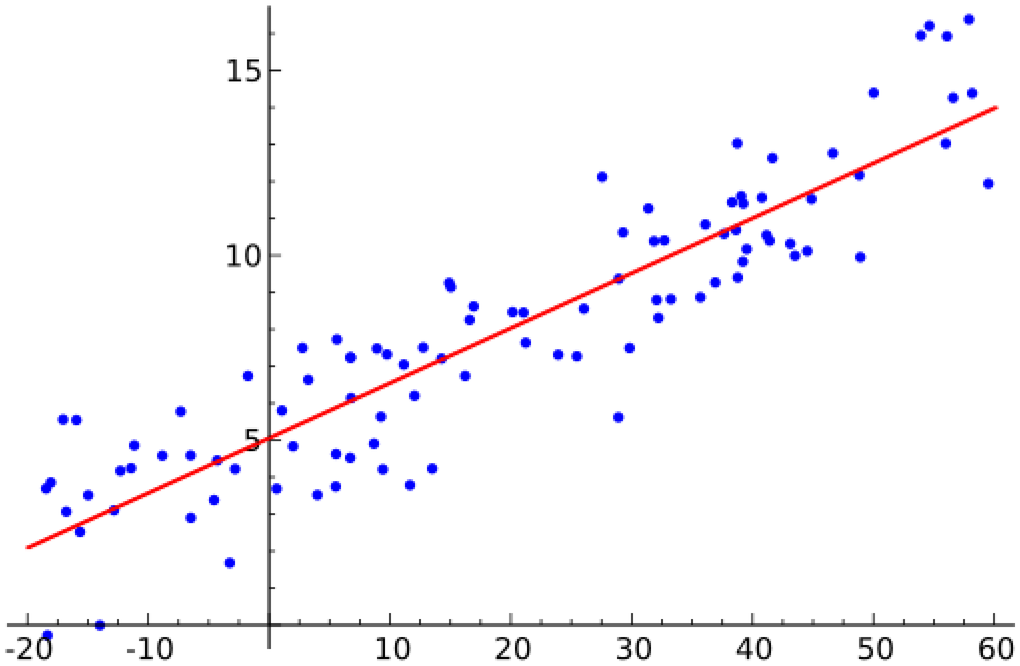
Prediction task

$$f : \mathcal{X} \to \mathcal{Y}$$

Model

- Class of functions where we are going to search
- Need a criteria: loss function and algorithm

Lineal regression



$$E_{in}(W) = \frac{1}{N}\sum_{n=1}^{N}(w^T x_n - y_n)^2$$

Gauss Markov theorem

Under the assumption of incorrelated noise, mean zero and bound variance, the Ordinary Least Squared technique reach the minimum variance unbiased estimator for β*

Model:
$y_i = f(x_i, \beta) + noise$, f linear in β

Minimizing E_{in}

$$\nabla E_{in}(w) = \frac{2}{N} X^T (Xw - y) = 0$$

$$X^T X w = X^T y$$

Result

$w = X^{\dagger} y$ where $X^{\dagger} = (X^T X)^{-1} X^T$

Gradient Descendent (Iterative method)

Given w_0 we want to find $\hat{v}$ such that $E_{in}(w_0 + \eta\hat{v}) < E_{in}(w_0)$

- Apply Taylor expansion to first order with $\|\hat{v}\| = 1$

$$\Delta E_{in} = E_{in}(w_0 + \eta\hat{v}) - E_{in}(w_0)$$

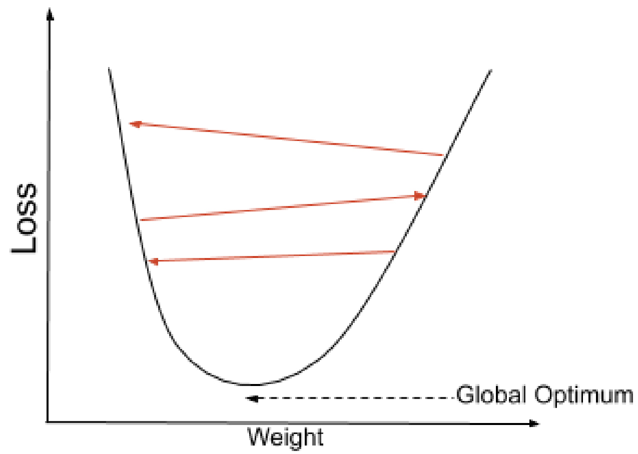
(...)

The equality holds if and only if

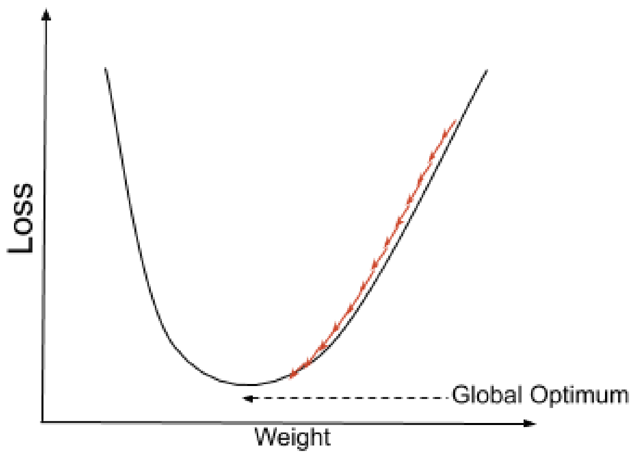
$$\hat{v} = -\frac{\nabla E_{in}(w(0))}{\|E_{in}(w(0))\|}$$

Negative Gradient! so reaches LOCAL optimun

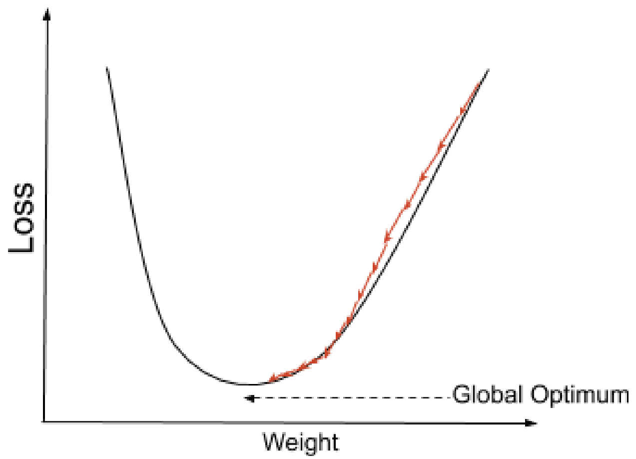
How η affects the algorithm



To High-Overshooting Problem



To-Low- Take Time to Train



Lower the learning rate as the training progresses

Learning rate

Perceptron (McCulloch- Pitts)

Neuronal Network

Is this a polynomial: more or less :)